Piscine datascience - 0

Creation of a DB

Summary: *Today, you will discover the creation of a DB*

Version: 1.00

Contents

Chapter I

General rules

- You have to render your modules from a computer in the cluster either using a virtual machine:

  - You can choose the operating system to use for your virtual machine

  - Your virtual machine must have all the necessary software to realize your project. This software must be configured and installed.

- Or you can use the computer directly in case the tools are available.

  - Make sure you have the space on your session to install what you need for all the modules (use the goinfre if your campus has it)

  - You must have everything installed before the evaluations

- Your functions should not quit unexpectedly (segmentation fault, bus error, double free, etc) apart from undefined behaviors. If this happens, your project will be considered non functional and will receive a 0 during the evaluation.

- We encourage you to create test programs for your project even though this work **won't have to be submitted and won't be graded**. It will give you a chance to easily test your work and your peers' work. You will find those tests especially useful during your defence. Indeed, during defence, you are free to use your tests and/or the tests of the peer you are evaluating.

- Submit your work to your assigned git repository. Only the work in the git repository will be graded. If Deepthought is assigned to grade your work, it will be done after your peer-evaluations. If an error happens in any section of your work during Deepthought's grading, the evaluation will stop.

- By Odin, by Thor ! Use your brain !!!

Chapter II

Introduction

In the next two modules, you will see the role of a data engineer.

This second step is important to understand. The data engineer "cleans" the data and transforms it in order to have data ready to be analyzed by analysts/data scientists.

The next module involves data cleansing. This second step is important to understand the data engineer "cleans" the data and transforms it. The objective is to have data ready to be analyzed by analysts/data scientists.

We are at the end of February 2022, it's your first day in a company selling items on the Internet. Before leaving on a trip your boss gives you the sales of the last 4 months. You will have to exploit them and propose solutions to increase the turnover of the company.

Be careful with this "piscine". Even if you manage to validate a module, you may be stuck later if you haven't cleaned up or stored your data properly.

Chapter III

Exercise 00

	Exercise 00
	Exercice 00 : Create postgres DB
Turn-in directory : *ex00/*	
Files to turn in : `None`	
Allowed functions : `All`	

For this exercise you can use directly postgres if installed in your campus or if you go through a VM, otherwise you have to use docker compose.

- The username is your student login

- The name of the DB is piscineds

- The password is "mysecretpassword"

We must be able to connect to your posgress database with this command:

```
psql -U your_login -d piscineds -h localhost -W
mysecretpassword
piscineds=#
```

Chapter IV

Exercise 01

	Exercise 01
Exercice 01 : Show me your DB	
Turn-in directory : *ex*01/	
Files to turn in :	
Allowed functions : `pgadmin, Postico, dbeaver or what you want to see the db easily`	

- Find a way to visualize the db easily with a software

- The chosen software must help you to easily find and manipulate data using its own corresponding ID

Chapter V

Exercise 02

	Exercise 02
	Exercice 02 : First table
Turn-in directory : *ex02/*	
Files to turn in : `table.*`	
Allowed functions : `All`	

- Create a postgres table using the data from a CSV from the 'customer' folder. Name the tables according to the CSV's name but without the file extension, for example : "data_2022_oct"

- The name of the columns must be the same as the one in the CSV files and have the appropriate type, beware you should have at least 6 different data types

- A DATETIME as the first column is mandatory

> Be careful, the typings are not quite the same as under Maria DB

Chapter VI

Exercise 03

	Exercise 03
	Exercice 03 : automatic table
Turn-in directory : *ex03/*	
Files to turn in : `automatic_table.*`	
Allowed functions : `All`	

- We are at the end of February 2022, you should be able to create tables with data extracted from a CSV.

- Now, in addition, retrieve all the CSV from the 'customer' folder automatically and name the tables according to the CSV's name but without the file extension, for example : "data_2022_oct"

Below is an example of the expected directory structure:

```
$> ls -alR
total XX
drwxrwxr-x 2 eagle eagle 4096 Fev 42 20:42 .
drwxrwxr-x 5 eagle eagle 4096 Fev 42 20:42 ..
drwxrwxr-x 2 eagle eagle 4096 Jan 42 20:42 customer
drwxrwxr-x 2 eagle eagle 4096 Jan 42 20:42 items


./customer:
total XX
drwxrwxr-x 2 eagle eagle 4096 Fev 42 20:42 .
drwxrwxr-x 5 eagle eagle 4096 Fev 42 20:42 ..
-rw-rw-r-- 1 eagle eagle XXXX Mar 42 20:42 data_2022_dec.csv
-rw-rw-r-- 1 eagle eagle XXXX Mar 42 20:42 data_2022_nov.csv
-rw-rw-r-- 1 eagle eagle XXXX Mar 42 20:42 data_2022_oct.csv
-rw-rw-r-- 1 eagle eagle XXXX Mar 42 20:42 data_2023_jan.csv


./items:
...
```

Chapter VII

Exercise 04

	Exercise 04
	Exercice 04 : items table

Turn-in directory : *ex04/*

Files to turn in : `items_table.*`

Allowed functions : `All`

- You have to create the table "items" with the same columns as in the "item.csv" file

- You have to create at least 3 data types in the table

Below is an example of the expected directory structure:

```
$> ls -alR
total XX
drwxrwxr-x 2 eagle eagle 4096 Fev 42 20:42 .
drwxrwxr-x 5 eagle eagle 4096 Fev 42 20:42 ..
drwxrwxr-x 2 eagle eagle 4096 Jan 42 20:42 customer
drwxrwxr-x 2 eagle eagle 4096 Jan 42 20:42 items

./customer:
...

./items:
total XX
drwxrwxr-x 2 eagle eagle 4096 Fev 42 20:42 .
drwxrwxr-x 5 eagle eagle 4096 Fev 42 20:42 ..
-rw-rw-r-- 1 eagle eagle XXXX Mar 42 20:42 items.csv
```

Chapter VIII

Submission and peer-evaluation

Turn in your assignment in your `Git` repository as usual. Only the work inside your repository will be evaluated during the defense. Don't hesitate to double check the names of your folders and files to ensure they are correct.

> The evaluation process will happen on the computer of the evaluated group.